September 11, 2013

Improving the Dependability of Research in Personality and Social Psychology: Recommendations for Research and Educational Practice

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. Personality and Social Psychology Review, 18(1), 3-12. (IR #18441)

Running head: Improving the dependability of research

Corresponding author:
David Funder
Department of Psychology -075
University of California, Riverside
900 University Avenue
Riverside, CA 92506
USA
funder@ucr.edu

UU IR Author Manuscript

Abstract

The Society for Personality and Social Psychology (SPSP) Presidential Task Force on Publication and Research Practices was appointed in response to concerns about the dependability and replicability of research findings in personality and social psychology, a problem that also plagues fields as diverse as physics, economics, biochemistry, medicine and cell biology. In this article the Task Force offers a brief statistical primer and recommendations for improving the dependability of scientific research. Recommendations for research practice include (1) describing and addressing the choice of *N* (sample size) and consequent issues of statistical power, (2) reporting effect sizes and 95% confidence intervals for findings, (3) avoiding "questionable research practices" that can undermine the assumptions underlying statistical procedures and inflate the probability of Type I error, (4) making available research materials necessary to replicate reported results, (5) adhering to SPSP's data sharing policy, (6) encouraging publication of high quality replication studies, and (7), maintaining flexibility and openness to alternative standards and methods when evaluating research. Recommendations for educational practice include (1) encouraging a culture of "getting it right" rather than "finding significant results," (2) teaching and encouraging transparency of data reporting, (3) improving methodological instruction on topics such as effect size, confidence intervals, statistical power, replication, and the effects of questionable research practices, and (4) modeling sound science and supporting junior researchers who seek to "get it right." The hope is that these recommendations can help lead the way to improved research practices and a more transparent research culture, throughout all of science.

UU IR Author Manuscript

Improving the Dependability of Research in Personality and Social Psychology: Recommendations for Research and Educational Practice

The SPSP Presidential Task Force on Publication and Research Practices was appointed in February 2013 and charged with making recommendations to the Executive Committee concerning actions to improve the dependability and replicability of research findings in personality and social psychology. The impetus for this task force arose in response to growing concerns about the dependability and replicability of research findings in fields as diverse as physics, economics, biochemistry, medicine, and cell biology, as well as in psychology. As a behavioral science organization, SPSP is well positioned to be a leader in improving research practices and professional communication across disciplines. The hope is that the actions taken by SPSP will be a model for other organizations within and outside of psychology.

The current wave of concern about the dependability of research findings arguably started with Ioannidis's (2005) provocatively titled paper, "Why most published research findings are false." The paper raised issues that apply to much of science, arguing that incentive structures and research practices produce a high rate of false positive findings. In psychology, the paper's shockwaves were amplified by a series of events including the publication of Vul et al.'s (2009) critique of social and affective neuroscience, followed by reactions – ranging from incredulous to disdainful – to Bem's (2011) article on extrasensory perception. Shortly thereafter, a number of prominent symposia and articles criticized research practices in psychology said to be widespread (e.g., Simmons, Nelson, & Simonsohn, 2011; John, Loewenstein, & Prelec, 2012). Around the same time, and presumably coincidentally, three well-known social/personality psychologists each retracted multiple papers, and in one case admitted to falsifying results, while other cases of data fraud emerged in fields including biology, oncology, genetics and even dentistry.

It is important to note that the research practices questioned by Vul, Simmons, John and others (and illuminated in the specific case of the Bem article) have nothing to do with data fraud. The criticisms addressed how studies are designed, analyzed and reported -- including some

UU IR Author Manuscript

practices that have become almost traditional -- and did not question the basic integrity of anyone's data. However, the contemporaneous emergence of cases of data fraud drew added attention to already-regnant concerns about the dependability of published research findings.

Critiques of psychology's research methods and practices are far from new (e.g., Carver, 1978; Cohen, 1994; Meehl, 1967). However, the conflation of recent events has led to calls for reform that are unprecedented in breadth and intensity. And, by and large, the field of psychology is responding (perhaps more so than many other scientific disciplines; Yong, 2012). The Association for Psychological Science announced several major initiatives aimed at improving the dependability of research published in their journals, the Psychonomic Society revamped publication guidelines, and other journals, organizations, and even government agencies are in the midst of similar examinations. Things are changing, and as the largest organization for social/personality researchers in the world, SPSP is in a unique position and has a special obligation to take a leading role in shaping these changes.

In April 2013, the SPSP task force met to generate recommendations for steps to improve the quality of research practices and the dependability of research findings. The purpose of this article is to outline these recommendations and some of the basic principles and statistical issues that lie behind them.[1] Although the need for the task force arose due to the events of the past few years, our recommendations are forward-looking. Their broad goal is to improve the quality of research, so that psychology can do an even better job in pursuing its core mission of understanding human beings as the complex social creatures they are.

How can "research quality" be defined? Many suggestions focus on replicability. Although replication may be the ultimate test of a scientific hypothesis or theory, it is not the only indicator of "truth-value," and many worthwhile studies are difficult or impossible to replicate (e.g., the effects of unique events such as the September 11, 2001 terrorist attacks on emotions and social behavior). Other indicators such as statistical power, precision of estimate, reliability, and internal, construct, and external validity are also important and deserve at least as much

[1] These recommendations included specific suggestions for actions by the publications, training, and awards committees of SPSP; the full report to the EC is available on-line at www.spsp.org.

UU IR Author Manuscript

attention (Shadish, Cook, & Campbell, 2002). Thus, the task force recommendations consider all of these criteria, as well as seeking to make research practices more transparent and to improve researchers' education and training. To set the stage for these recommendations, we review a few basic but essential statistical concepts.

A Brief Statistical Primer

Social and personality psychologists work hard to design informative research. Yet, the researcher has no control over many aspects of a study, including the participant's genetics, early environment, and even what happened to the participant on the way to the laboratory. Hence, findings of a study based on one sample of 100 participants cannot be expected to be exactly the same as those based on another sample of 100 participants drawn from the same population. This state of affairs gives rise to the need for statistics to allow researchers to address variation, often substantial in magnitude, that occurs from sample to sample (sampling error) and to make an inference about what would happen if they could measure the entire population. The statistical concepts that underlie such inference have important implications for how research should be conducted, analyzed and reported.

Type 1 Error: The Type 1 error rate (α) is the *conditional* probability that the present data, or even more extreme data, will be observed in a given sample *given* a specific condition in the population. As typically used in social and personality psychology, that condition is that the relationship (treatment effect; correlation) in the population is precisely 0. A relationship of 0 in the population means that any non-zero results observed in a specific sample are purely due to chance. Following a suggestion by Sir Ronald Fisher, a convention that the Type 1 error rate should not exceed .05 has been widely accepted. The goal of this convention is to set an acceptable upper limit on the likelihood that findings will be reported as "significant" when there actually is no relationship in the population from which the current sample was drawn.

This technique of "null hypothesis significance testing" (NHST) continues to be widely used despite criticisms expressed repeatedly over the years (e.g., Carver, 1978; Fraley & Marks,

UU IR Author Manuscript

2007). As Jacob Cohen (1994, p. 997) noted, "What we want to know is 'Given these data, what is the probability that H_0 [the null hypothesis] is true?' But as most of us know, what the obtained *p*-value tells us is 'Given that H_0 is true, what is the probability of these (or more extreme) data?'" Another common criticism is that, because the obtained *p*-value varies with *N* (the number of participants or independent observations included in a particular study), it is not a measure of the magnitude of the finding -- although it is often erroneously interpreted as such.

Effect Size: The magnitude of the statistical relationship found, the "effect size," may be expressed in unstandardized (raw) units or in standardized (*z*-score) units. For an experiment on the effects of priming on response time, the mean unstandardized difference in response time (*Y*) between primed and unprimed treatment conditions might be $\overline{Y}_{prime} - \overline{Y}_{control}$ = 5.0 milliseconds. For a correlational study on the relationship between fathers' and sons' heights, the unstandardized regression coefficient (slope) for predicting the height of an adult son from his father's height might be B_1 = 0.5 inches (i.e., for each 1 inch increase in the height of the father, the son will be on average 0.5 inches taller). In areas with clear consensus that the measurement units are at least interval level (e.g., seconds, mm blood pressure), unstandardized effect sizes are preferred. However, in many areas of social and personality psychology, such well-developed measurement units are not available (e.g., 5-point Likert type ratings; different scales used across studies). In these cases, standardized effect sizes are advisable. For experiments comparing treatment (*t*) and control (*c*) conditions, a commonly used standardized effect size is Cohen's $d = \frac{\overline{Y}_t - \overline{Y}_c}{SD}$, which divides the difference between the means of the treatment and control groups by an estimate of the standard deviation of the difference (*SD*). For correlational relations, standardized effect sizes include the Pearson correlation *r* and the squared partial correlation (pr^2). We focus below on the most commonly used effect size statistics, *d* and *r*, which are estimates of their corresponding parameter values in the population, δ and ρ.

Perhaps because *d* is traditionally used in experimental research and *r* is more often used in correlational research, it is not universally understood that these effect sizes have a

mathematical relationship. The Pearson *r* can be converted to Cohen's *d* via the formula $d = \frac{2r}{\sqrt{[1-r^2]}}$, and the reverse conversion is $r = \frac{d^2}{\sqrt{d^2+4}}$ (Rosenthal, 1991).

Statistical Power: When the null hypothesis of no relationship is rejected, the conventions of NHST lead to the conclusion that an alternative directional hypothesis (e.g., the response time in the prime condition is less than in the control condition) can be accepted.

Statistical power is the *conditional* probability that a true effect of a precisely specified size (e.g., δ = 0.5 or ρ = 0.3) in the population will be detected in a study using such conventional significance testing. Statistical power is 1 – the Type 2 error rate, where the Type 2 error rate is the conditional probability that a *true* effect of the precisely specified size will *not* be detected under NHST. Recall that statistical power and the Type 1 error rate are conditional probabilities; each only applies when the relevant condition is met. Type 1 error applies when the effect is 0 in the population; statistical power applies when the precisely specified (e.g., δ = 0.5 or ρ = 0.3) effect size characterizes the population.

An important goal in designing research is to maximize statistical power, the probability that the null hypothesis will be rejected if there is, in fact, a true effect of the specified size in the population. However, this goal can be challenging -- statistical power will be limited by factors such as sample size, measurement error, and the homogeneity of the participants. Cohen (1988) suggested a convention that investigations should normally have power = 0.8 to detect a true effect of the specified size in the population. This value assumes a Type 2 error rate (.20) that is four times the conventional Type 1 error rate (.05). Other, typically higher, values of power (e.g., 0.9) have been suggested in other disciplines (e.g., Lenth, 2001) and for certain types of investigations (e.g., important replication studies; tests of important applied programs).

Relations amongType 1 Error, Standardized Effect Size, Statistical Power, and Sample Size: The exact observed *p*-level, standardized effect size, and sample size are mathematically inter-

UU IR Author Manuscript

related (Rosenthal, 1991). If any two are known, the third can be directly computed.[2] This basic point is not universally understood. For example, claims that researchers only need to care about the *p*-level of an effect but not its size evaporate in the light of the recognition that to report or to base a decision on one of these numbers is precisely equivalent to reporting or basing a decision on the other, given a particular *N*.[3] For illustration, imagine that a researcher finds the result that $p = .05$ (two tailed) in a study with $N = 80$. The estimate of the standardized effect size will then be $r = .22$. Given this relationship, the conventional practice of setting $\alpha = .05$ (two tailed) as the critical threshold for significance in this study is precisely equivalent to employing the standard that the result should not be reported if the effect size estimate is lower than $r = .22$.

Despite this equivalence, focusing solely on the observed *p*-level is problematic because findings with equivalent *p*-levels can have very different implications. In cases where *N*s are very large (e.g., when working with census data), extremely small effect sizes may achieve statistical significance, yet carry no important theoretical or practical meaning. In other cases, where *N*s are very small (e.g., in experiments with only a handful of subjects per condition), significant effects may imply implausibly large effect sizes. An implausibly large effect size, especially when paired with a small *N*, may be an irreproducible outlier ("fluke" finding) or even, in rare cases, a leading indicator of improper research practices.

The problems with focusing exclusively on the observed *p*-level are exacerbated when researchers over-rely on the dichotomous distinction between "significant" and "non-significant" results. This common practice risks treating nearly equivalent findings as if they were importantly different, especially if one finding barely attains the $p < .05$ threshold whereas

[2] For example, in meta-analytic practice, a standard method for estimating effect sizes is to convert the reported exact *p*-level to *t*, and then employ the formula, $r = \sqrt{\frac{t^2}{t^2+(n_1+n_1-2)}}$, where n_1 and n_2 are the sizes of the two samples (or experimental groups) being compared (Lipsey & Wilson, 2001). This conversion, and others, can also be performed using online calculators such as http://www.campbellcollaboration.org/resources/effect_size_input.php.

[3] Cumming (2012) notes the extreme variability in exact *p*-values that can result even in identical replications of the same experiment simply due to natural variation from sample to sample, a fact little appreciated by many researchers. He illustrates this variability in an entertaining and informative brief video http://www.youtube.com/watch?v=ez4DgdurRPg.

UU IR Author Manuscript

the other barely misses it. We have seen cases in which researchers reported a significant and theoretically expected effect, then reassured readers by showing that a further effect that would disconfirm their theory (or reveal a confound) did not achieve significance. Yet the confidence intervals of the two effect sizes showed substantial overlap![4] If the difference between two effects is theoretically important, then they should be tested for whether they are significantly different from each other, not just from zero. More generally, the routine reporting of effect sizes and confidence intervals would help prevent researchers from drawing misleading interpretations of this sort.

The relationship among *p*-value, effect size, and sample size extends directly to statistical power. Once any three of the four statistical quantities are known, the fourth can be easily calculated with freely downloadable, user friendly software like G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) and MBESS (Kelley, 2007). Two common analysis scenarios involve these quantities. These analyses should be carried out *before*[5] the investigation is conducted because they provide important information regarding whether the planned study has sufficient statistical power to be able to detect the hypothesized effects of interest.

In the first scenario, an α (typically .05), a value of statistical power (typically 0.8), and a range of reasonable estimates of the standardized effect sizes in the population are considered in order to determine what sample size is needed. Estimates of standardized effect sizes may be gleaned from (in order of preference) meta-analyses, prior research, pilot studies, or norms established by Cohen (1988) or by investigators in the substantive area of research for small, moderate, and large effect sizes. Many researchers in social psychology are unaware of the

[4] In other words, although the first effect is "significant" and the second is not, the two effects are not necessarily significantly different from each other.

[5] Statisticians strongly advocate a priori power calculations. Less known is that observed (post hoc) power analyses sometimes suggested by behavioral science researchers are *not* recommended by statisticians (e.g., Hoenig & Heisey, 2001; Lenth, 2001). Observed power calculations typically do not provide the desired information and can lead to nonsensical conclusions in some applications. Confidence intervals or equivalence tests described later in this article are more likely to provide the desired information. Yuan and Maxwell (2005) show analytically and through simulation that the results of post hoc power analyses are biased and often associated with large errors of estimation. They conclude that when the estimate of observed power is low, "the observed power may not provide any useful information regardless of the sample size!" (p. 163). Observed power analyses ignore the confidence interval associated with the observed effect size and the nonlinear relationship between effect size and statistical power. As noted by Lenth (2007), "Researchers owe it to themselves to take a thoroughly prospective view of any power calculation" (Lenth, p. 11).

UU IR Author Manuscript

sample sizes required to achieve adequate statistical power. For example, if an experiment on priming were to assign an equal number of participants to the priming and control conditions ($n_{prime} = n_{control}$), a total of 788 participants (394 in each condition) would be needed to detect a small standardized effect size ($\delta = 0.2$ standard deviation difference), 128 participants would be needed to detect a moderate standardized effect size ($\delta = 0.5$), and 52 participants would be needed to detect a large effect size ($\delta = 0.8$) with power = 0.8. These values indicate that experiments with 10 or 20 participants per condition – which are not uncommon – are seriously underpowered except in the case of large effect sizes that are rare in personality and social psychology. Statistical studies of the published literature in clinical, personality, and social psychology journals have found that the typical investigation in these fields has a statistical power of approximately .45 to .65 to detect a moderate effect size in the population ($\delta = 0.50$ or $\rho = .30$); some other areas (e.g., health psychology) typically exceed power = .80 (see Rossi, 2013).

In the second scenario, an α, a feasible sample size (e.g., 100 participants), and reasonable estimate(s) of the standardized effect size are chosen in order to calculate power. If an experiment includes 100 total participants, 50 in each treatment group, the estimated statistical power will be 0.17 to detect a small, .70 to detect a moderate, and .98 to detect a large effect size. Methods also exist for increasing statistical power without increasing *N* (see Dennis, Lennox, & Foss, 1997; Shadish et al., 2002, Table 2.3); these methods involve procedures (e.g., more powerful treatments, more reliable measurement, more homogeneous participants, more adequate treatment of missing data) that increase the standardized effect size.

A meta-analysis of a wide range of social psychological phenomena found an overall average published effect size of $r = .21$ (or $d = .46$) (Richard, Bond, & Stokes-Zoota, 2003). A smaller meta-analysis of personality research found exactly the same average published effect size, $r = .21$ (Fraley & Marks, 2007). While perhaps an overestimate because of publication bias, this effect size provides one plausible candidate for benchmarking the effect size estimate that could be used for the calculation of statistical power.

UU IR Author Manuscript

Recommendations for Research Practice

The task force recommends several "best practices" for research in personality and social psychology, most of which are based on the statistical concepts and their relationships summarized above. While not intended as hard-and-fast rules (see Recommendation 7, below), we believe that these recommendations are sufficiently important that researchers should take them into account when planning, analyzing and reporting their research in SPSP journals or elsewhere.

Recommendation 1. *Describe and address choice of N and consequent issues of statistical power.*

Researchers should design studies with sufficient power to detect the key effects of interest. Often, research will involve multiple types of effects (e.g., effects of treatments on the key outcome; mediational analyses) that can be expected to have different effect sizes. The sample size should normally be justified based on the smallest effect of interest. For example, consider a 2 x 2 design in which a hypothesized main (average) effect of treatment is expected to be large in magnitude (δ = 0.8) and a theoretically equally important treatment x gender interaction is expected to be moderate in magnitude (δ = 0.5). The researcher should base the power calculation on the magnitude of the smaller treatment x gender interaction effect. This difference can be consequential: The *n* needed to achieve .80 power increases from approximately 52 total participants for the *d* = 0.8 main effect to approximately 128 total participants for the *d* = 0.5 interaction effect, assuming participants have been equally divided among the four groups.

We recognize that research involving small populations (e.g., rare diseases), time-intensive methods (e.g., coding naturalistic behavior), longitudinal data gathered over extended periods of time, or larger units of analysis (e.g., group dynamics research) may not be able to achieve high levels of statistical power. For some newer statistical procedures, no known mathematical solution for calculating statistical power may yet exist. (In these cases, several statistical packages [e.g., Mplus, see Muthén & Muthén, 2002 for an introduction] now include relatively easy to use simulation routines that provide good empirical approximations of statistical

power.) It is important to underline that studies that do not achieve high statistical power should not be dismissed out of hand; they can still gather data that are informative and worthy of publication. Nonetheless, we recommend that *a priori* statistical power be reported whenever possible and considered as one factor among many when interpreting results. One potential salutary effect of conducting more studies with adequate power will be that more statistical effects of interest will achieve conventional (.05) levels of statistical significance. Consequently, any temptation to engage in questionable data analytic practices in order to achieve these conventional levels of significance is likely to be lower.

Recommendation 2. *Report effect sizes and 95% confidence intervals for reported findings.*

Even though they are related, observed p-values, effect sizes, and confidence intervals highlight different aspects of the results, so they provide complementary information. *p*-values need to be supplemented by effect sizes that provide information on the magnitude of a finding. Effect sizes provide a clear metric for the comparison of results across studies (is the present result large/small relative to prior research?), and form the basis for meta-analyses summarizing entire bodies of research. As noted earlier, if there is good agreement on the units of the effect (e.g., reaction time in milliseconds; weight in kilograms), unstandardized effect sizes are preferred. When no such agreement exists, standardized effect sizes should be reported. The task force recommends that either unstandardized or standardized effect sizes be reported, as appropriate. Occasionally this will not be possible because methods for calculating effect sizes for newly proposed advanced statistical procedures do not yet exist.

Confidence intervals add an assessment of the precision of the estimate to the effect size measure. For the typical two-group between-subjects experiment in which different participants receive the treatment and control conditions, the confidence interval (CI) represents the difference between the group means ± a margin of error associated with sampling variability. The CI = $\left(\overline{Y}_t - \overline{Y}_c\right) \pm t_{critical}\left(se_{\overline{Y}_t - \overline{Y}_c}\right)$ where $t_{critical}$ is the (tabled) value of the *t*-distribution corresponding to the level of Type 1 error rate selected (typically $\alpha = .05$) and the degrees of freedom ($df = n_t + n_c - 2$), where $se_{\overline{Y}_t - \overline{Y}_c}$ is the standard error of the difference between the means. If the confidence interval overlaps with 0, it also means that the possibility

UU IR Author Manuscript

of no relationship in the population is plausible, and therefore the null hypothesis, H0: $\mu_1 - \mu_2 = 0$, cannot be rejected. A 95% confidence interval means that the true effect will be included in the confidence interval 95% of the time across repeated investigations using samples of the same size from the same population. As was mentioned earlier, when effects within a study are theoretically expected to be different from each other, examination of the CIs illustrates whether this expectation was met (Cumming & Finch, 2005). By the same token, examination of the overlap between the confidence intervals of multiple replication studies provides far more information about the conclusions that should be drawn than simple examination of the obtained *p*-values or whether the results of each study were statistically significant or not (Cumming, 2012).

For the earlier priming experiment example, a 95% confidence interval of 4.5 to 5.5 milliseconds implies a far more precise estimate of the difference in the mean reaction time between the primed and control conditions than a 95% confidence interval of 0.01 to 9.99 milliseconds. The latter finding reflects substantial imprecision in the results and should therefore stimulate appropriate caution in drawing theoretical inferences. Moreover, while confidence intervals that do not include zero offer conventional grounds for rejecting the null hypothesis, such intervals may still include ranges of non-zero effect sizes that are too small to be theoretically informative.

Evaluation of effect size is a matter for scientific interpretation because small effects sizes can be potentially be important. Moreover, standardized effect sizes can be affected by the strength of the experimental manipulation, the precision of measurement, the homogeneity of the sample, whether the research was conducted in the laboratory or in the field, and many other factors. The definition and interpretation of the scientific importance of small and large effect sizes, therefore, depends on the nature of the research question, the research context, and the substantive domain.

Confidence intervals can be easily constructed for most types of effects, but sometimes complications arise. Some confidence intervals (e.g., for the Pearson *r*) are not symmetric and require a normalizing transformation (e.g., the Fisher *r* to *z* transformation); others do not have

UU IR Author Manuscript

a known mathematical solution and can only be constructed empirically through repeated sampling procedures (e.g., bootstrapping). Researchers can compute a priori estimates of the sample size needed to produce a confidence interval with a desired width with user-friendly software (see Kelley, 2007; Kelley & Maxwell, 2012). Increasing sample size and improving the reliability of dependent measures can both help achieve tighter confidence intervals. The task force recommends that a conventional 95% confidence interval be reported to provide an estimate of both the size and the precision of the effect.

Recommendation 3: *Avoid "questionable research practices."*

Recommendations 1 and 2 assume that the researcher has proposed hypotheses prior to conducting a study, has tested hypotheses "appropriately," and has reported findings "fully." Differences of opinion can certainly exist about what constitute "appropriate" analytic strategies and "full" reporting of results in a particular study. Nonetheless, procedures that look at the results and then tweak the data post hoc to achieve statistical significance undermine the ability of researchers to reach a valid conclusion about the existence of an effect in the population. In a review of articles published in three journals, Masicampo and Lalande (2012) found that *p*-values between .045 and .050 were reported far more often than would be expected statistically. When data have been tweaked, the reported effect size will almost certainly be far greater in magnitude than the true effect size—which might well be 0. Such practices greatly increase the likelihood that even a well-designed replication study will fail to support the original finding.

Therefore, the following research practices are widely regarded as questionable: (1) conducting multiple tests of significance on a data set without statistical correction; (2) running participants until significant results are obtained (i.e., data-peeking to determine the stopping point for data collection); (3) dropping observations, measures, items, experimental conditions, or participants after looking at the effects on the outcomes of interest; and (4) running multiple experiments with similar procedures and only reporting those yielding significant results. These practices may not be equally problematic; both (3) and (4) have particularly great potential to

UU IR Author Manuscript

lead to serious inflation of the Type 1 error rate and yet not be recognized in the review process (Simmons et al., 2011).

We fully acknowledge that in the context of exploratory research and certain other cases some of these practices may be justifiable and even wise. When problems of interpretation arise, it is often because of how studies and analyses were reported, not how they were conducted. Therefore, if researchers feel that these research practices are warranted for a given study, great care should be taken to fully describe the research and analytic process. The findings should be clearly described as exploratory to avoid representing tentative discoveries as conclusive findings until they are replicated or otherwise verified (Diaconis, 1985).

Recommendation 4. *Include in an appendix the verbatim wording (translated if necessary) of all independent and dependent variable instructions, manipulations and measures. If the manuscript is published, this appendix can be made available as an on-line supplement to the article.*

Researchers wishing to replicate an existing study or to conduct a new study that builds on earlier research need to know the precise procedures of the prior research. Historically, space limitations have precluded complete reporting of the details of studies, necessitating extensive correspondence with the author to fully ascertain key procedures. This limitation is no longer relevant given web-based storage. The increased transparency provided by the availability of the full description of the study will provide readers with a greater ability to evaluate the results of the study and to conduct related research.

Recommendation 5. *Adhere to SPSP's "Data Sharing Policy" which states that: "The corresponding author of every empirically-based publication is responsible for providing the raw data and related coding information underlying all findings reported in the paper to other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that a) the confidentiality of the participants can be protected; b) legal rights concerning proprietary data do not preclude their release; and c) those requesting data agree in writing in advance that shared data are to be used only for the purpose of verifying the substantive claims through reanalysis or for some*

UU IR Author Manuscript

other agreed-upon use. " Adopted by the Executive Committee of the Society for Personality and Social Psychology, July 19, 2013.

Open access to data is the norm in most scientific disciplines once results based on those data have been published. Many of the US funding agencies also require sharing of publicly funded data after investigators have published their findings. At the same time, participant confidentiality (e.g., the possibility that any participant could be uniquely identified from information in the data base) and legal agreements concerning proprietary data (e.g., data made available by an organization to a researcher) must be honored. Unlike some areas of science, psychology has not yet developed a norm of verifying important findings through reanalyses of data. This norm may be reflected in the relative paucity of erratum reports in our journals. Reanalyses not only can identify errors made by the original authors, but can reveal heretofore unconsidered features of the data that clarify the theoretical contribution of the study. We encourage researchers to document and archive the dataset on which their reported analyses are based at the time they submit the original research report. This practice facilitates easy access of the data if they are requested at some future point in time.

Recommendation 6. *Encourage, and improve the availability of publication outlets for replication studies.*

Many researchers would agree that replicability is the *sine qua non* of scientific knowledge. There is much to value in the "scientist's belief in 'stubborn facts' with a life span that is greater than the fluctuating theories with which one tries to explain them" (Shadish, et al., 2002, p. 31). Yet replication studies traditionally have been difficult to fund and to publish. Funding agencies prioritize new and “transformative” research topics, and many journals implicitly – and sometimes explicitly – discourage the publication of replication studies. However, establishing a firm foundation upon which findings can accumulate will be useful in ultimately pushing research forward, by helping researchers to avoid premature closure and blind alleys. Some settled research questions may not be as settled as commonly believed. Further, a great deal of wasted research time and resources might be avoided if researchers could be more confident in the published literature.

UU IR Author Manuscript

Therefore, we suggest that funding agencies reserve some proportion of their resources for high quality replication studies, either as independent projects (if warranted by the importance of the research question) or as part of research programs exploring new topics. We also believe that the journal that originally published a prominent finding has a special obligation to publish high quality research that replicates or fails to replicate that finding, rather than automatically consigning such research to lower-visibility or limited "replication" journals. At the same time, replication studies should be evaluated against the high standard of quality enforced for the journal as a whole. In the case of replication research, hallmarks of high quality include adequate power (and the more the better, perhaps suggesting a benchmark of .90 or .95 for adequate power for single replication studies rather than the conventional .80), multiple studies, sound methodology, and high theoretical importance. These are all matters for editorial judgment. We would further suggest that the research community allocate higher value to replication research than it has traditionally received in the past. We particularly encourage research that integrates replication studies into progressive, creative research programs that have the potential to contribute to both the underlying foundation of "stubborn facts" as well as to make innovative contributions to knowledge.

Recommendation 7. *Maintain flexibility and openness to alternative standards and methods when evaluating research.*

Notwithstanding everything said above, we do not advocate inflexible rules. One of the hallmarks of the scientific peer-review process is that each paper is evaluated individually, in the context of its specific subfield, and according to ever-improving data analytic techniques. Standards of evaluation should shift across studies and over time, and editors and reviewers should be flexible. Some research requires special populations, methods, or data analyses, making it impossible to apply the same standards across the board. Any reform movement risks going too far – imposing new standards so strictly that the diversity of research questions and methods is stifled. One of the strengths of social/personality psychology, perhaps what puts our discipline at the heart of the field of psychology (Yang & Chiu, 2009), is its methodological

UU IR Author Manuscript

diversity. We should balance consistently rigorous standards with attention to the unique challenges of different research questions and methodologies.

Recommendations for Educational Practice

The recommended research practices would, we believe, increase the quality of published research. However, there is the chance that those who adopt these practices before they become commonplace may be at a disadvantage in the publication and hiring/promotion process. Thus, in order to make our field more amenable to these practices, it is important for all of us, including editors, reviewers, and those who make hiring/promotion decisions, to educate ourselves about their value.

Recommendation 1: *Encourage a culture of "getting it right" rather than "finding significant results."*

The beginning of this educational process is to encourage a culture of "getting it right" (accurate knowledge) over "successful" studies (valuing only predicted statistically significant effects) (Asendorpf et al., 2013). Venues for encouraging this culture include those with an explicit educational purpose, such as graduate and undergraduate courses, textbooks, workshops, and methodological articles, as well as those where the educational purpose is more implicit, such as editorial guidelines and instructions to reviewers and grant panels. All of these venues provide opportunities for teaching or reminding both experienced and novice researchers that the contribution of a particular piece of research should be evaluated in terms of whether the research is carefully designed to address important and interesting questions, whether care has gone into operationalization and measurement, and whether the characteristics of the particular sample (both its nature and size) are appropriate to the questions being asked and the generalizability of the conclusions that are drawn. In addition, it is critical that the statistical analyses used are appropriate for the questions and the nature of the data collected. If these criteria are met, then the research is likely to be valuable whether or

UU IR Author Manuscript

not (all) the results come out as expected, or analyses clearly identified as exploratory address questions that arose only after unexpected initial results.

Recommendation 2: *Teach and encourage transparency of data reporting, including "imperfect" results.*

Researchers sometimes feel under pressure to conduct studies that can be completed quickly, to adjust their hypotheses to fit their results, and/or to provide incomplete information about their methodology or findings if things “did not work”. The recommendations put forward here encourage, instead, a focus on the informativeness of data despite occasional messiness. Although omitting non-significant or unexpected findings can help the flow of a paper, it is important to keep that information available somewhere, if not in the paper then in supplemental materials available to readers. In our various roles-- as mentors of students, as authors, editors, reviewers, and grant panel members—we need to promote a climate that emphasizes "telling the whole story" rather than "telling a good story."

Recommendation 3: *Improve methodological instruction on topics such as effect size, confidence intervals, statistical power, meta-analysis, replication, and the effects of questionable research practices.*

We encourage graduate and undergraduate courses in statistics, research methods and ethics, as well as workshops and tutorials open to those at all stages of their careers, to include training about the issues raised in these recommendations. These include the consequences of using questionable research practices and the usefulness of effect sizes, confidence intervals, and statistical power. Students should also learn about the importance of meta-analytic thinking, why replication is important, and the unique challenges of replication research.

Recommendation 4: *Model sound science and support junior researchers who seek to “get it right.”*

The burden for improving the field should not fall mainly on new researchers. Perhaps the best way to effect change is to model improved research practice and alter the incentive structures from the top down. Established researchers can do two things. First, they can demonstrate

UU IR Author Manuscript

proper research practices by conducting and publishing sound science, correctly analyzed and transparently reported, which may entail following practices different from those commonly used in the past. Second, and perhaps even more importantly, they can encourage and support the publication, hiring, and promotion of junior researchers who put "getting it right" ahead of "publishing significant findings." Sometimes, a shorter vita may be a better one.

Some Reflections on the Implications of Statistical Power for Replication Studies

The late meta-analyst John Hunter wryly offered his observations on the progress of research in many areas of psychology given that researchers often ignore considerations of effect size and statistical power. According to Hunter, a research area begins with the proposal of an interesting hypothesis and the excitement of a first demonstration study that finds a large effect size. Subsequent research tries to clarify the phenomenon by designing studies to rule out alternative explanations, thereby making the effect size smaller. This stage is followed by a generation of studies investigating mediation and moderation, which further reduce the effect size. Researchers continue to use informal guidelines for sample size gleaned from the experience of the initial demonstration study. The result is that replication of the original effect becomes less and less common due to decreased statistical power. Finally comes the inevitable review paper: “Where is the (insert name) effect?” Different stages of research will be associated with different effect sizes. Careful attention to effect size, sample size, and statistical power is thus needed as research progresses.

Moreover, reviewers and editors must expect a less than perfect match of results across multiple studies, multiple measures, and multiple analyses. If four independent exact replications of a study are properly conducted with a statistical power of 0.8 to detect the true effect size in the population, the probability that all of the replications will be statistically significant is only 0.4! If the power is lower than .80 -- as is common in personality and social psychology-- the probability of an unbroken series of significant replications is even lower, perhaps to the point of implausibility (Schimmack, 2012). Rather than expecting uniformly significant effects across studies, careful examination of the match of the hypothesized pattern

UU IR Author Manuscript

to the obtained pattern of effect sizes across measures and studies, as advocated by Donald Campbell, can help reduce this problem. Meta-analyses of the results across multiple studies can provide even better estimates of the mean and variability of the effect sizes.

The methodological issues in conducting a replication study can be challenging, with statistical power deserving special consideration. Often the original effect size can be estimated from only a single study (or a few studies). Maxwell (2013) and Dallow and Fina (2011) remind us that the true population effect size is not known; rather the effect size is estimated so that it has a confidence interval. In addition, the relation between effect size and power is not linear, so the effects on statistical power when the effect size estimate is too low are not mitigated by cases in which the effect size estimate is too high. Finally, we must be very cautious in concluding that a study did not replicate and therefore there is no effect. This conclusion implies the questionable practice of accepting the null hypothesis that the effect size is precisely 0. Maxwell's analysis implies that researchers seeking to replicate the results of a study may want (a) to consider using a lower than reported effect size to calculate power when it is estimated based on a single study, (b) to develop a minimum value for an effect size that is deemed too small to be of interest, so that a test of nonequivalence (Rogers, Howard, & Vessey, 1982; Seeman & Serlin, 1998) can be performed testing more definitively whether the obtained effect is less or greater than that value, and (c) to use a higher value than .80 for statistical power to enhance the probability that the replication clarifies rather than further confuses the finding. Each of these steps enhances the credibility and usefulness of the replication study. Alternative Bayesian statistical approaches such as calculating the Bayes factor (Kass & Raftery, 1995) can also be informative about the relative likelihood that each competing hypothesis is true.

Conclusion

Donald Campbell (see Overman, 1988) long espoused his belief in the capacity of mutual criticism to improve scientific practice and ultimately promote understanding of the "truth" of scientific claims. The development of a field of personality and social psychology that values replication is a step towards the culture he envisioned. But, for such a culture to thrive, it is

UU IR Author Manuscript

important that responses to replication studies should be civil and focus solely on issues of research methodology and substantive theory. Failures by others to replicate one's work should be treated as opportunities to work together with colleagues to find the parameters under which a theoretically expected effect is and is not found. Critiques of research methodology or empirical findings merit constructive, not defensive, responses. By the same token, critiques and replication studies should be undertaken as open-minded investigations of the generalizability of important, interesting effects, not as cynical attempts to "score points" or undermine established findings. As psychologists we know that such guidelines can be challenging to follow. But, to the extent we can focus on critical scientific issues and foster a culture of "getting it right," our field will enjoy more rapid scientific progress as well as better collegial relations.

Finally, while some of the recommendations offered in this article address particular aspects of psychological research, most of them – especially those regarding the promotion and acceptance of replication studies – address issues that are common to many areas of research in the physical, life and behavioral sciences (e.g., Rehman, 2013). Our hope is that these recommendations can help lead the way to improved research practices and a more transparent research culture, throughout all of science.

References

Abelson, R.P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129–133.

Asendorpf, J.B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J.J.A., Fiedler, K., Fiedler, S., Funder, D.C., Kliegl, R., Nosek, B.A., Perugini, M., Roberts, B.W., Schmitt, M., Vanaken, M.A.G., Weber, H., & Wicherts, J.M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108-119.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407-425. doi:http://dx.doi.org/10.1037/a0021524

UU IR Author Manuscript

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378-399. Retrieved from http://search.proquest.com/docview/616376408?accountid=14521

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round, p < .05. *American Psychologist, 49,* 997-1003.

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *6*0, 170-180.

Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics*, *10*, 311-317.

Dennis, M. L., Lennox, R. D., & Foss, M. (1997). Practical power analysis for substance abuse health services research. In K. J. Bryant, M. Windle, & S.G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 367-404). Washington, DC: American Psychological Association.

Diaconis, P. (1985). *Theories of Data Analysis: From Magical Thinking Through Classical Statistics*. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring Data Tables, Trends and Shapes* (1-36). New York: Wiley.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology (pp. 149-169)*. New York: Guilford.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*, 19-24.

UU IR Author Manuscript

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS: Medicine, 2*(8): e124. doi:10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi:http://dx.doi.org/10.1177/0956797611430953

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*, 970-984.

Kelley, K. & Maxwell, S. E. (2012). Sample size planning. In H. Cooper (Ed.) *APA handbook of research methods in psychology* (181--202). Washington, DC: American Psychological Association.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician*, *55*, 187-193.

Lenth, R. V. (2007). Post hoc power: Tables and commentary. Technical report No. 378, Department of Statistics and Actuarial Science, University of Iowa. Retrieved September 8, 2013 from http://www.stat.uiowa.edu/sites/default/files/techrep/tr378.pdf

Maxwell, S. E. (2013, August). *Methodological issues in planning and interpreting replication studies*. Invited address, American Psychological Association, Honolulu, HI.

Lipsey, M.W., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology*, *65*, 2271-2279.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103-115.

Muthén, L.K. and Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *4*, 599-620.

Overman, E. S. (1988). *Methodology and epistemology for social science: Selected papers of Donald T. Campbell*. Chicago: University of Chicago Press.

UU IR Author Manuscript

Rehman, J. (2013, Sept 1). Cancer research in crisis: Are the drugs we count on based on bad sciences? *Salon.* http://www.salon.com/2013/09/01/is_cancer_research_facing_a_crisis/.

Richard, F.D., Bond, C.F., Jr., & Stokes-Zoota, J.J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331-363.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553-565.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Revised edition). Newbury Park, CA: Sage.

Rossi, J.S. (2013). Statistical power analysis. In J.A. Schinka & W.F. Velicer (Eds.), *Handbook of psychology. Volume 2: Research methods in psychology* (2nd ed., pp. 71–108). New York: Wiley.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17,* 551-566.

Seaman, M.A., & Serlin, R.C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3,* 403-411.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. doi:http://dx.doi.org/10.1177/0956797611417632

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274-290. doi:http://dx.doi.org/10.1111/j.1745-6924.2009.01125.x

Yang, Y., & Chiu, C. (2009). Mapping the structure and dynamics of psychological knowledge: Forty years of APA journal citations (1970–2009). *Review of General Psychology, 13*(4), 349-356. doi:http://dx.doi.org/10.1037/a0017195

Yuan, K.-H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, *30*, 141-167.

Yong, E. (2012, May 17). Replication studies: Bad copy. *Nature, 485,* 298-300.